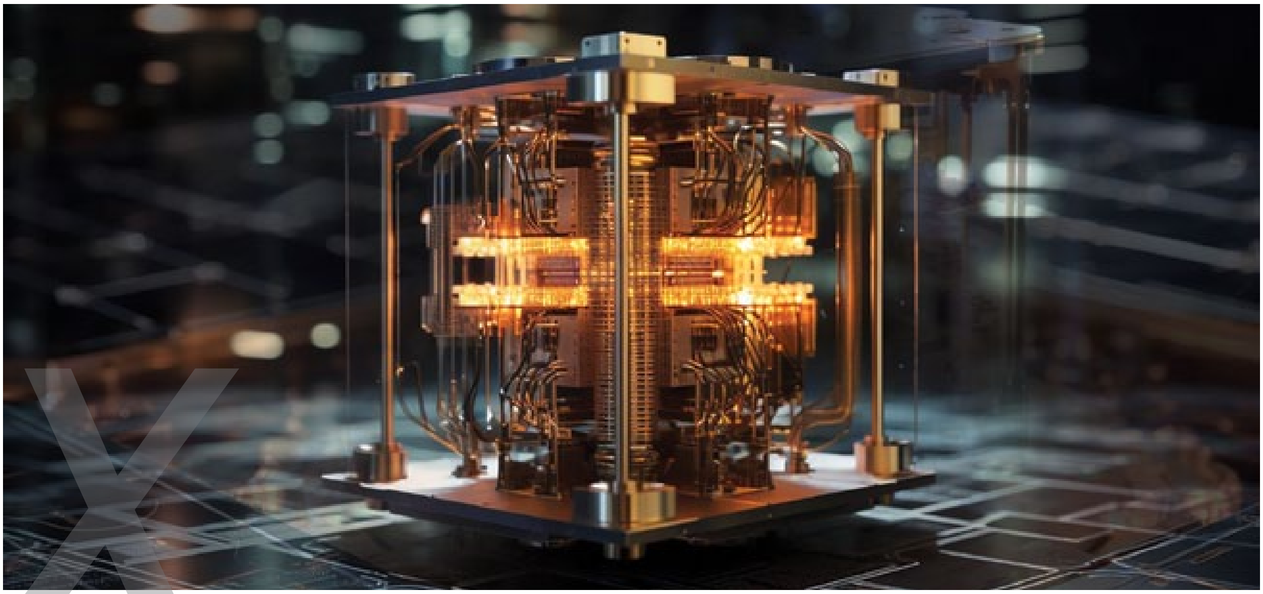
# Supercomputing on Ojammpic!

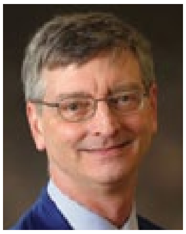
## Would Quantum breakthroughs and AI Supercomputers finally make the elusive super-machine affordable, accessible and lean enough to sit happy next to your coffee mug?

By Pratima H



XS! Two golden alphabets that sound like honey to anyone getting out of a gym and into a shopping aisle. Being able to shrink something that was L to S is never final but always fulfilling. From huge halls housing clunky clusters of metal beasts to mainframes skimmed into client-servers which, too, shrunk into PCs that, later, collapsed themselves to fit ultra-thin smartphones—computing has kept getting leaner, lighter and closer to the Average Joe and Jane. As also predicted very cryptically by all the laws circling the worlds of software and hardware. Things will shrink. And when they do, they will be an arm-length away.

Supercomputers could not have asked for a better moment than now to follow suit. There's quantum computing rubbing its eyes and asking for its morning beans. There's AI shouting from the bench press to flex those chips, tone those engines and stretch towards those extra flops. After what we have heard recently (Google's Willow to Nvidia's $3000 personal AI supercomputer DIGITS) – looks like things are going to get more trimmed ahead? Does that mean the 'super' in computing is all set to get 'skinny'? Would that mean fitting well into the sizes of affordability, democratisation and efficiency? Or would that mean



> The NAIRR pilot at our campus has democratisation of access to AI resources as a specific goal, and we are both a major partner in implementing this with NSF and we provide significant resources to the NAIRR pilot.
>
> **- Bill Gropp,** Illinois Grainger Engineering

> Increasing synergy between cloud and edge, decreasing costs of HPC will support supercomputers become more accessible and widespread.
>
> **- Mukesh Ranjan**
> Everest Group

something counter-intuitive? Like 'The Rock' wearing a school-girl's pretty T-shirt? Would crunched-up turn into crammed-up?

### TIGHTER VICTORIAN CORSETS

Things are moving. Fast. AI Hyperscaler CoreWeave has announced plans about shipping Nvidia GB200 Grace Blackwell Superchip-enabled AI supercomputers to IBM. Nvidia has unwrapped the Nvidia Jetson Orin Nano Super Developer Kit- something that, it says fits in the palm of a hand; and will help commercial AI developers, hobbyists and students to tap generative AI capabilities and performance from their desks. The specs already read: a 1.7x leap in generative AI inference performance, a 70 per cent increase in performance to 67 INT8 TOPS. And, of course, the big showstopper at CES: Project DIGITS, a 'personal AI supercomputer' with access to its Grace Blackwell hardware platform in a compact form factor. And it might be able to run models up to 200 billion parameters in size. We also heard Oracle talking about how its Oracle Cloud Infrastructure (OCI) is taking orders for the largest AI supercomputer in the cloud—available with up to 131,072 Nvidia Blackwell GPUs. The chops: 2.4 zettaFLOPS of peak performance. Plus claims of max levels of three times as many GPUs as the Frontier supercomputer and more than six times that of other hyperscalers.

There were also reports of an AI training supercomputer. Condor Galaxy, a network of nine interconnected AI supercomputers, popping up with the handshake of the US-based AI company Cerebras Systems together with G42, UAE-based technology holding group G42. Read it as a 4 exaFLOPs, 54 million core, cloud-based AI supercomputer that can dial up to 36 exaFLOPs of AI computing. (For reference: One exaFLOP is the ability for a supercomputer to perform one quintillion (1018) floating-point operations per second or FLOPs). Other giants are also joining this race to the 'mini' aisle. Expect AWS to flesh out Project Rainier, the supercomputer, an UltraCluster baked with its Trainium chips, which are intended for AI programs. Google is already out with Willow – a quantum chip that can reduce errors exponentially as more qubits kick in. Apparently, Willow performed a standard benchmark computation in less than five minutes- something that would take one of today's fastest supercomputers 10 septillion (1025) years. As Google said in its blog: "Willow brings us closer to running practical, commercially relevant algorithms that can't be replicated on conventional computers."

Supercomputing used to be an exclusive domain of governments and research institutions, as it required vast budgets and specialized expertise, reflects Devroop Dhar, Co-Founder and Board Member at Primus Partners. "However, this landscape has been reshaped with breakthroughs in semiconductor technology, cloud-based High-Performance-Computing (HPC) and Artificial Intelligence (AI). These emerging trends aim to make supercomputing accessible to a wider audience."

"AI supercomputers and desktop supercomputers are rapidly becoming mainstream realities. The demand for AI-specific workloads like large language models, autonomous vehicles and real-time analytics is driving this innovation at much faster pace," adds Dhar. "Companies and governments worldwide are investing in AI supercomputers like OpenAI's Azure infrastructure, Japan's Fugaku and many more, to meet these computational needs. At the same time, advances in hardware have enabled compact, affordable systems capable of supercomputing-level performance, making desktop supercomputers viable for researchers, developers and small businesses." As these trends mature, the distinction between traditional supercomputers and consumer-level systems will blur and this will make supercomputing power an integral part of industries and even personal computing.

But wait! Are trends like democratisation, miniaturisation, commodification and consumerisation not counterintuitive to supercomputing, per se?

### THE TUX WORN AS A T-SHIRT

Counter-intuitive? No. Asserts Louiqa Raschid, Dean's Professor of Information Systems in the Smith School with a joint appointment with the Institute for

Supercomputing is not a monolithic concept anymore, it's is going to be an adaptable, accessible resource.

**- Devroop Dhar**
Primus Partners

Advanced Computer Studies and the Department of Computer Science at the University of Maryland.

"These trends are not counterintuitive. The use of ML/AI pipelines and large-scale data analytics is becoming more commonplace and non-specialist users will need access to computing platforms that are more affordable and potentially provide access to less specialised compute nodes," Prof. Raschid explains.

Supercomputing is fundamentally about scope (specific use cases like drug discovery, weather forecasts etc.) through scale, highlights Dr. Nityesh Bhatt is Professor and Chairperson of Information Management Area, Institute of Management, Nirma University. "Initiatives like DIGITS have potential to enhance this scope by allowing many more people and companies to innovate with almost similar scale but at an affordable price." He argues that this will supplement supercomputing (not counterintuitive). "With project DIGITS, vision of Nvidia is to place an AI supercomputer on the desks of every data scientist, AI researcher and student and to empower them to engage and shape the age of AI."

Mukesh Ranjan, Vice President, Everest Group also does not see these trends as counterintuitive, but a natural evolution to advancements in supercomputing technology.

"Trends such growing AI/ML demand, hardware advancements, increasing synergy between cloud and Edge, decreasing costs of high-performance computing, and evolution within developer and consumer ecosystem will support supercomputers become more accessible and widespread."

We ask the same question to Bill Gropp, Distinguished Chair in Engineering at Illinois Grainger Engineering, University of Illinois Urbana-Champaign and he answers, "This is an interesting question, though naturally with a complex answer. There are several ways to look at it."

Prof. Gropp who is also the Director of the National Center for Supercomputing Applications first points out how today's smartphone is far more capable than supercomputers of a few decades ago (i.e., when I was a grad student). "So in that sense, personal supercomputers are here now. One very old definition of a supercomputer was 'the computer that cost US$15 million.' Of course, today that is US$200 million+. But by that definition, there are always problems so big and so important that you need the biggest machine that you can afford. In that sense, supercomputers will always be rare and available only to a few."

His assessment beckons us to zoom in on the word 'affordable' once more.

### WHERE'S THE COMMON-STREET CRINOLINE?

The triage of AI, quantum and supercomputing is going to be unprecedented and shake up things in all three worlds. But would we get democratisation and affordability when this dust falls? Supercomputers are not exactly going to be milk-cartons that we can pick from a nearby store. Not too soon, at least.

It's clear that AI use-cases are driving a lot of demand for HPC and supercomputing capabilities, reckons Balaji Padmanabhan, Director, Center for Artificial Intelligence in Business and Professor of Decision, Operations and Information Technologies for the University of Maryland's Robert H. Smith School of Business. "AI models are compute–and data-hungry for training, and compute-hungry for "inference" (i.e. when the AI models are used, such as querying ChatGPT). Training will likely be a non-stop process of continually improving models, which require such HPC capabilities. But every time AI models are used, some billion-parameter neural network is being 'fired' to produce tokens–an expensive process that runs mostly on such HPC infrastructures in the cloud."

Dhar examines why or why not these new trends spell access. "Democratisation ensures that businesses of all sizes and even individuals can access computational power through cloud platforms like AWS, Azure and Google Cloud. Miniaturization has led to compact yet powerful systems, such as Nvidia's AI-specific GPUs or Apple's M-series chips, that deliver near-supercomputing performance in desktop or

> Every time AI models are used, some billion-parameter neural network is being “fired” to produce tokens - an expensive process that runs mostly on such HPC infrastructures in the cloud.
>
> **- Balaji Padmanabhan**
> Center for Artificial Intelligence in Business, University of Maryland

portable devices. Commodification and consumerisation help to further reduce costs and simplify usage. This is helping to bring supercomputing capabilities closer to everyday applications.”

But Prof. Gropp also adds another dimension here. “I think the sense of the question is a third point: AI gives us a way to do some computations that previously required a supercomputer (of the big and expensive type), and in this sense, AI has greatly accelerated the rate at which highly capably computing is widely available. So in many (but not all!) cases, what had until recently only been possible with a supercomputer is now available far more broadly, driven by a combination of the advances in computing hardware and software and AI.” He cites here the example of computational fluid dynamics, such as flow around a truck or aircraft (or inside an air conditioning duct). “We know how to compute the flow but doing it for a large-scale object is hard and computationally intensive, especially when the flow becomes turbulent. AI approaches can make it easier to computing those flows, though it requires a conventional supercomputer to help train the AI model.

There is one more question that cannot be ignored, as Dr. Bhatt reminds. “Not many professionals require this much scale and advanced configuration. At a starting price of US$3000 ( 2.6 lakh), end-users cannot think of buying it, specifically in the developing countries. Additional cost may be incurred for add-ons and other support necessitating additional funding support. It may lead to some form of rental model or EMI-based buying.”

I think we are still a while away from true democratisation in terms of everyone having access to enough computing resources to build their own powerful AI models (there are some recent promising trends though in this direction), contends Prof. Padmanabhan. “The best scenarios today are users having access to some limited capabilities—often expensive—to fine-tune some of the foundation large models built by the big tech companies (on their expansive computing hardware). I don’t see that trend changing anytime soon given how expensive AI hardware is. But the advent of powerful AI infrastructure on the cloud, usually owned by large companies and in some cases, by government or educational institutions, does make this available to anyone (who has the resources to pay for the use) to build their own AI models.”

Prof. Gropp gives us a peek at some of the work going on at his university campus.

“We’re doing many things like—research in using AI to replace computations previously requiring a supercomputer. The fluid flow example is just one of many. Also—Providing AI-optimized computing for researchers. There is still a role for supercomputers, especially for AI training and research into AI methods. But these systems can be made more available and easier to use that “classic” supercomputers. We’re doing this as part of several NSF-funded projects. Our Delta and DeltaAI systems provide computing to researchers across the country. The NSF ACCESS program helps connect researchers with our systems and with others at other institutions (we lead several parts of ACCESS and are major participants in others).”

### THE AI LYCRA IS HERE, REST OF THE SPANDEX, ON THE WAY

Pocket-friendly or Pocket-sized—in many ways, Computing is, at least, not as rigid, off-hands and enormously expensive as it used to be.

“We are indeed seeing trends towards some of the chip innovation focused on Edge devices, like mobile phones or tablets. Apple, Nvidia and others have hardware today that makes it possible to run AI models locally on small devices, which is enabling use-cases where individuals may interact with language models locally as they compose or read documents. Making this easier is also a trend toward ‘small language models’ that are optimised for some narrow task, enabling such local computing possible,” observes Prof. Padmanabhan.

Hardware, historically, has seen trends where increasing power eventually makes it to desktops,

> I expect that there will be a smaller segment of users that can justify the costs associated with a dedicated localised supercomputer.
>
> **- Louiqa Raschid**
> Robert H. Smith School of Business, University of Maryland

and while true 'desktop supercomputers' are still a while away, there will be significantly more powerful desktop models that can run complex AI models locally without incurring costly computing costs through inference on the cloud, Prof. Padmanabhan weighs in. "Nvidia has announced the new "Grace Blackwell Superchip" which will permit desktops to become fairly powerful AI machines that can be used to build (and use) fairly large AI models locally. This trend will continue making it possible for desktop machines to have enough compute capabilities for AI prototyping. This is similar in a sense to data scientists originally having to use cloud resources to train and run machine learning models for prediction on large tabular data. Over time many of these use-cases were possible to do locally, given how powerful desktops became."

As to whether the likes of AI supercomputers and desktop supercomputers be a mainstream reality soon, Prof. Raschid cites how HPC platforms to support AI pipelines are already available. "Some are customised and tuned for application domains, e.g., FinTech pipelines. Desktop supercomputers may be a more niche item since I expect that there will be a smaller segment of users that can justify the costs associated with a dedicated localised supercomputer."

These will not necessarily mean mainstream adoption, at least over the next five to seven years, augurs Ranjan. "AI supercomputers still remain too expensive for general use, have high energy requirement, pose sustainability challenges, and have very specialised use-cases where RoI can be justified. In the near term, we can expect adoption in specialised industries such as scientific and medical research, aerospace and defense, and hi-tech, among others. Gradually, we can expect trickle-down adoption within enterprises as costs decreases and demand grows for AI-driven applications."

The way Dhar dissects it, these new trends do not undermine traditional supercomputers but complement their evolution. "They expand the utility of supercomputing beyond niche scientific and governmental domains to industries like healthcare, climate modelling, media, autonomous systems and many more. While supercomputers remain essential for large-scale simulations and data processing, smaller, democratised systems are opening up new opportunities for even small size businesses.

Software vendors can surely leverage it for augmenting their AI-powered services for the clients, contemplates Dr. Bhatt. "I am optimistic that this will pave the way for similar innovations from other technology vendors in future which will lead to further price reduction."

### THE SIZE UK VS. SIZE US DILEMMA, AGAIN

Ultimately, it's the 80-20 rule in a sense at work here, as Prof. Padmanabhan quips. "Such compute scenarios will not solve all problems (e.g. such as updating a search engine's index for instance, or training personalisation models for billions of consumers), but it might solve 80 per cent of use-cases that most organisations have in terms of AI and ML. With access to such hardware, it does make sense to do some work locally given benefits in cost (i.e. not having to pay continuous inference costs) and latency."

True access means a lot of things to a lot of people. For developers, it is all about the ability to play with new AI possibilities from their desks. For enterprises iffy about investing in AI (without the burden of AI computing and infrastructure), it could mean something else. For academic campuses and research tanks, the word 'accessible' would surely wear different clothes. In computing, it is important to shrink but without shrivelling or shoe horning. Sometimes it is about FLOPs. But sometimes it is about dollars, or seconds or a use-case. What fits one may be tight or loose to someone else. Same for the wallet.

So let's hope that supercomputers will continue this path without turning into the iron-bed that Procrustes used. True XS – and not some tortured or force-fitted XXL – is what the world needs here. Something that just does not fit. But hugs.

*pratimah@cybermedia.co.in*